ON THE RANDOM SAMPLING OF PAIRS, WITH PEDESTRIAN EXAMPLES

RICHARD ARRATIA AND STEPHEN DESALVO

Abstract. Suppose one desires to randomly sample a pair of objects such as socks, hoping to get a matching pair. Even in the simplest situation for sampling, which is sampling *with* replacement, the innocent phrase "the distribution of the color of a matching pair" is ambiguous. One interpretation is that we condition on the event of getting a match between two random socks; this corresponds to sampling two at a time, over and over without memory, until a matching pair is found. A second interpretation is to sample sequentially, one at a time, with memory, until the same color has been seen twice.

We study the difference between these two methods. The input is a discrete probability distribution on colors, describing what happens when one sock is sampled. There are two derived distributions — the pair-color distributions under the two methods of getting a match. The output, a number we call the *discrepancy* of the input distribution, is the total variation distance between the two derived distributions.

It is easy to determine when the two pair-color distributions come out equal, that is, to determine which distributions have discrepancy zero, but hard to determine the largest possible discrepancy. We find the exact extreme for the case of two colors, by analyzing the roots of a fifth degree polynomial in one variable. We find the exact extreme for the case of three colors, by analyzing the 49 roots of a variety spanned by two seventh-degree polynomials in two variables. We give a plausible conjecture for the general situation of a finite number of colors, and give an exact computation of a constant which is a plausible candidate for the supremum of the discrepancy over all discrete probability distributions.

We briefly consider the more difficult case where the objects to be matched into pairs are of two different kinds, such as male-female or left-right.

Contents

Date: November 27, 2012.

1. Motivation

The problem that inspires us: Suppose a drawer has 12 white and 4 black socks. How many socks must one remove to ensure a pair of matching color? The answer, 3, illustrates the pigeon-hole principle. The statement of detailed counts, 12 and 4, was arbitrary, but leads to the problem that we address in this paper: what is the distribution of the color of a matching pair?

To simplify, we take the limit as the number of socks in the drawer goes to infinity, while the proportions remain constant, e.g., seventy five percent white and twenty five percent black.

We consider two sensible methods for choosing "a matching pair."

(M1) Select objects two at a time until a pair of the same color is selected in a single round;

(M2) Select objects one at a time until the first pair of the same color is found.

For a second example, if there are 365 equally likely colors for socks, then, under Method 2 the maximum number of socks inspected is 366, but the expected number is $23.6166\ldots$.[1] In contrast, the expected number of pairs inspected by Method 1 is exactly 365, hence the expected number of socks inspected is 730. However, our focus is not on the *number* of socks inspected, but rather, on the distribution of the *color* of the matching pair.

[1]The exact computation is $\mathbb{E}\,N = \sum_{k\geq 0} \mathbb{P}(N > k) = \sum_{k=0}^{364} (365)_k/365^k$, with the notation $(n)_k = n!/(n-k)!$ for n falling k.

In our first example, under method 1 the odds for a white pair over a black pair are $(12/16)^2$ to $(4/16)^2$; equivalently 12^2 to 4^2, or 3^2 to 1^2, so that 9/10 of the time the pair is white, and 1/10 of the time it is black. Under method 2, the outcomes resulting in a white pair correspond to ww, bww, wbw, with total probability $(.75)^2 + 2(.75)^2(.25)^2 = 27/32$, and the outcomes resulting in a black pair correspond to bb, wbb, bwb, with total probability $(.25)^2 + 2(.75)(.25)^2 = 5/32$.

To summarize, the input is a distribution on colors, $\mathbf{p} = (.75, .25)$, and there are two outputs: under Method 1, the color of a pair is white with probability .9, and black with probability .1, while under Method 2, color of a pair is white with probability 27/32, and black with probability 5/32.

$$\begin{aligned} \mathbf{p} &= (.75, .25) \\ M1(\mathbf{p}) &= (.9, .1) \\ M2(\mathbf{p}) &= (.84375, .15625). \end{aligned}$$

Some natural questions, for an arbitrary discrete distribution $\mathbf{p}$ for the color of a single sock:

(Q1) When does $M1(\mathbf{p}) = M2(\mathbf{p})$?

(Q2) How far apart can $M1(\mathbf{p})$ and $M2(\mathbf{p})$ be from each other?

There are practical algorithms [1] for sampling, exploiting the birthday paradox, that require getting a matching pair whose color has the distribution (M1), but under a naive *opportunistic* implementation, would only find a pair whose color is distributed according to (M2). Question (Q2) above is about quantifying the error that would result from using the opportunistic implementation.

2. Pair-derived distributions

In general, we write S for the random color of a single sock, and describe the initial distribution of colors with

$$p_i := \mathbb{P}(S = i).$$

When the number of colors is finite, say $n+1$, then we let the colors be $0, 1, 2, \ldots, n$, and the distribution of S is given by $\mathbf{p} = (p_0, p_1, \ldots, p_n)$. Our initial example had $n + 1 = 2$, $\mathbf{p} = (p_0, p_1) = (.75, .25)$. When the number of colors is infinite, we take the colors to be $0, 1, 2, \ldots$, and then $\mathbf{p} = (p_0, p_1, p_2, \ldots)$.

Method 1 may be described as the color X of a pair of randomly chosen socks, conditional on getting a match. More precisely, the two

chosen socks have colors S and S' and are independent and identically distributed, with $\mathbb{P}_i = \mathbb{P}(S = i)$. We write

$$f_2 := \mathbb{P}(S = S') = \sum_i \mathbb{P}(S = S' = i) = \sum_i p_i^2 \tag{1}$$

for the probability that two randomly chosen socks match, so

$$\mathbb{P}(X = i) = \mathbb{P}(S = i | S = S') = \frac{p_i^2}{f_2}. \tag{2}$$

Method 2 involves a sequential procedure: pick socks one at a time until a duplicate color is found. Suppose that when this duplicate is found, there have been k *other* colors, with $k = 0, 1, 2, \ldots$. Write i for the duplicate color, and $J = \{j_1, \ldots, j_k\}$ for the single colors, so that $i \notin J$ and $|J| = k$. The second occurrence of color i is at time $k + 2$, and for the first $k+1$ socks, any permutation of the colors in $\{i\} \cup J$ is valid. Hence the color Y of the matching pair found by Method 2 has distribution given by

$$\mathbb{P}(Y = i) = p_i^2 \sum_k (k+1)! \sum_J p_{j_1} \ldots p_{j_k}. \tag{3}$$

In the sum above, $|J| = k$ and $i \notin J$.

3. When are the two pair-picking methods the same?

A discrete distribution is said to be *uniform* if it has finite support, say of size $n+1$, and for each color i in the support, $p_i = 1/(n+1)$. It is easy to see that if $\mathbf{p}$ is uniform, then $M1(\mathbf{p}) = M2(\mathbf{p})$.[2] The converse is true, but not so easy to prove; we will first prove an ancillary result in Lemma 1 and then summarize in Theorem 1.

Lemma 1. *Under Method 2, as specified by* (3)*,*

$$\text{if } p_i \geq p_j > 0, \text{ then } \frac{\mathbb{P}(Y = i)}{p_i^2} \leq \frac{\mathbb{P}(Y = j)}{p_j^2}, \tag{4}$$

hence

$$\text{if } p_i = p_j > 0 \text{ then } \mathbb{P}(Y = i) = \mathbb{P}(Y = j). \tag{5}$$

[2] Because, in fact, if $\mathbf{p}$ is a uniform distribution, then both $M1(\mathbf{p})$ and $M2(\mathbf{p})$ are equal to the original uniform distribution — by the principle of ignorance, all possible colors are alike, and hence, equally likely under each of the derived methods. We invite the reader to consider, is "principle of ignorance," i.e. invoking symmetry, without presenting details as in (5), an adequate proof?

Also,

$$\text{if } p_i > p_j > 0, \text{ then } \frac{\mathbb{P}(Y=i)}{p_i^2} < \frac{\mathbb{P}(Y=j)}{p_j^2}. \tag{6}$$

Proof. Assume $p_i \geq p_j > 0$. Define $t(i,k)$ to be the inner sum of (3), so that

$$\frac{\mathbb{P}(Y=i)}{p_i^2} = \sum_k (k+1)!\; t(i,k).$$

To prove (4) it suffices to show that if $p_i \geq p_j > 0$ then $t(i,k) \leq t(j,k)$ for all k, and to further prove (6), it suffices to show that if $p_i > p_j$ then $t(i,k) < t(j,k)$ for at least one k. With sums always taken over sets of size k,

$$t(i,k) = \sum_{i \notin J} p_{i_1} \cdots p_{i_k} = \sum_{i \notin J, j \in J} p_{i_1} \cdots p_{i_k} + \sum_{i,j \notin J} p_{i_1} \cdots p_{i_k},$$

that is, in the sum over sets J excluding i, we take cases according to whether or not $j \in J$. With a similar decomposition of $t(j,k)$, taking the difference yields

$$t(i,k) - t(j,k) = \sum_{i \notin J, j \in J} p_{i_1} \cdots p_{i_k} - \sum_{i \in J, j \notin J} p_{i_1} \cdots p_{i_k}.$$

There is a bijection between sets J for the first sum and sets J for the second sum, that substitutes i for j. From $p_i \geq p_j$ it follows that for all k, $t(i,k) \leq t(j,k)$, and further, when $p_i > p_j$, we have $t(i,1) < t(j,1)$. □

Theorem 1. *Over all discrete distributions* $\mathbf{p}$*, the derived distributions of* X *and* Y*, given by* (2) *and* (3)*, are equal if and only if* $\mathbf{p}$ *is a uniform distribution.*

Proof. Assume first that $\mathbf{p}$ is a uniform distribution, say over $n+1$ colors, so that for all i,j in the support of $\mathbf{p}$, we have $p_i = p_j = 1/(n+1)$. For i,j both in the support of $\mathbf{p}$, it is obvious from (2) that $p_i = p_j$ implies $\mathbb{P}(X=i) = \mathbb{P}(X=j)$, and (5) shows that $\mathbb{P}(Y=i) = \mathbb{P}(Y=j)$. Hence for i in the support of $\mathbf{p}$, $\mathbb{P}(X=i) = 1/(n+1) = \mathbb{P}(Y=i)$, implying that X and Y have the same distribution.

To prove the opposite direction, suppose $\mathbf{p}$ is *not* a uniform distribution. Then we can fix i,j with $p_i > p_j > 0$. From (6), we get

$$\frac{\mathbb{P}(Y=i)}{p_i^2} < \frac{\mathbb{P}(Y=j)}{p_j^2},$$

and dividing by f_2 to relate with (2), and rearranging,

$$\frac{\mathbb{P}(X=i)}{\mathbb{P}(X=j)} > \frac{\mathbb{P}(Y=i)}{\mathbb{P}(Y=j)}, \tag{7}$$

which implies that X and Y have different distributions. □

Theorem 1 gives a complete answer to our first question: when are the two pair-picking methods the same? Next we turn to the second question: when the two methods are different, how different can they be?

4. Total variation distance

We wish to quantify: given a probability distribution $\mathbf{p}$, with the matching pair chosen by Method 1 or Method 2, how far apart are the two distributions with respect to the color of the matching pair?

A natural metric on the space of all probability measures is the *total variation distance*.

Definition 1. *For two real-valued random variables X and Y, the total variation distance between the* laws *of X and Y is defined as follows.*

$$d_{\mathrm{TV}}(\mathcal{L}(X), \mathcal{L}(Y)) = \sup_{A \subseteq \mathbb{R}} |P(X \in A) - P(Y \in A)|,$$

where the sup is taken over all Borel sets $A \subseteq \mathbb{R}$.[3]

It is common to write $d_{\mathrm{TV}}(X, Y)$ instead of $d_{\mathrm{TV}}(\mathcal{L}(X), \mathcal{L}(Y))$.

Definition 2. *Given a discrete probability distribution $\mathbf{p}$, let X have the Method 1 distribution given by (2), let Y have the Method 2 distribution given by (3), and define the* discrepancy *of $\mathbf{p}$ by*

$$D(\mathbf{p}) = d_{\mathrm{TV}}(X(\mathbf{p}), Y(\mathbf{p})). \tag{8}$$

We could have written $D(\mathbf{p}) = d_{\mathrm{TV}}(X, Y)$ above, but we prefered $d_{\mathrm{TV}}(X(\mathbf{p}), Y(\mathbf{p}))$, to emphasize that $D(\mathbf{p})$ is the total variation distance between two probability laws, with each law being a function of a third underlying law $\mathbf{p}$.

Some elementary facts about total variation distance: When X and Y are discrete random variables, an equivalent definition is

$$d_{\mathrm{TV}}(X, Y) = \frac{1}{2} \sum_k |\mathbb{P}(X=k) - \mathbb{P}(Y=k)|, \tag{9}$$

[3]This choice of definition is useful for probability, with the desirable property that $d_{\mathrm{TV}}(X, Y) \leq 1$, and it equals $\sup_{f:\mathbb{R}\to[0,1]} |\mathbb{E}\, f(X) - \mathbb{E}\, f(Y)|$. But there is an alternate tradition, from analysis, to define the total variation distance between measures μ, ν as $\sup_{f:\mathbb{R}\to[-1,1]} |\int f d\mu - \int f d\nu|$, which, when applied to $\mu = \mathcal{L}(X), \nu = \mathcal{L}(Y)$, gives values ranging from 0 to 2.

and it follows, from $\sum_k \mathbb{P}(X=k) = \sum_k \mathbb{P}(Y=k)$, that by splitting up the summands into positive and negative parts,[4]

Lemma 2.

$$d_{\mathrm{TV}}(X,Y) = \sum_k (\mathbb{P}(X=k) - \mathbb{P}(Y=k))^+ = \sum_k (\mathbb{P}(X=k) - \mathbb{P}(Y=k))^- . \tag{10}$$

For example, when X is a Bernoulli random variable with parameter θ,[5] and Y is Bernoulli with parameter θ', the total variation distance is $|\theta - \theta'|$.

5. Special Cases

5.1. **Dimension $n = 1$: two colors of socks.** In the case $n = 1$, we write $\mathbf{p} = (p_0, p_1) = (x, 1-x)$. The discrepancy $D(\mathbf{p}) = d_{\mathrm{TV}}(X,Y)$ simplifies, via Lemma 2, to $|d_1|$, where

$$d_1(x) = \mathbb{P}(X=0) - \mathbb{P}(Y=0) = \frac{x^2}{x^2 + (1-x)^2} - (x^2 + 2(1-x)x^2).$$

The expression $|d_1(x)|$ is plotted in Figure 1.

Since d_1 is a rational function in one variable, it is easily optimized over $x \in [0,1]$. We outline our procedure as a preparation for the more difficult case in Section 5.2. We first put the derivative over a common denominator, which is strictly positive for $0 \le x \le 1$, and focus our attention on the numerator. The numerator is a sixth degree polynomial in x of the form $4\,(-x + 7x^2 - 18x^3 + 24x^4 - 18x^5 + 6x^6)$, having four real roots: 0, 1,

$$x_1 := \frac{1}{6}\left(3 + \sqrt{3\left(-3 + 2\sqrt{3}\right)}\right) \doteq 0.696660, \tag{11}$$

and the conjugate, $1 - x_1$. The list of roots already includes both endpoints of the domain $[0,1]$. The cusp for $|d_1(x)|$ at $x = 1/2$ is also critical, with $|d_1(1/2) = 0|$ corresponding to the uniform case. Evaluating $|d_1(x)|$ at these five critical numbers exhausts all possible extremes, and the maximum value is $d_1(x_1) = \frac{1}{\sqrt{135+78\sqrt{3}}} \doteq 0.0608468$.

[4]Notation: $t^+ = \max(0,t), t^- = \max(0,-t)$; hence $|t| = t^+ + t^-$ and $t = t^+ - t^-$.

[5]so that $\mathbb{P}(X=1) = \theta = 1 - \mathbb{P}(X=0)$

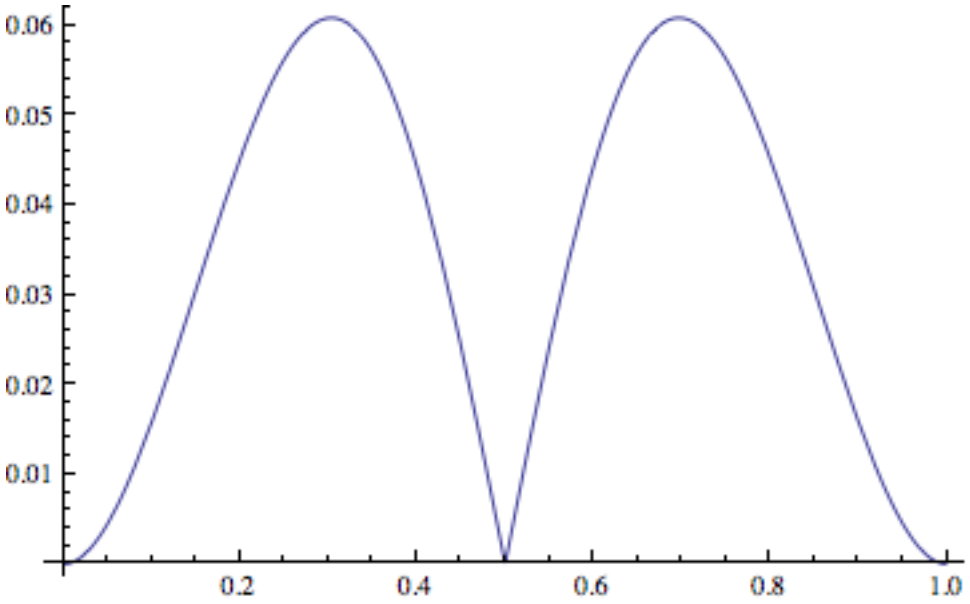

FIGURE 1. Plot of $D(\mathbf{p})$ for $\mathbf{p} = (x, 1-x)$, as a function of $x \in [0, 1]$.

5.2. **Dimension $n = 2$: three colors of socks.** The case $n = 2$ can be set up similarly to $n = 1$, but now we have three cases of possible signs underlying absolute values. Each case is a smooth, two-dimensional surface, and we find extremes by checking all critical values arising from points where the gradient vanishes, and on the boundary. To avoid subscripts, we switch notation from $\mathbf{p} = (p_0, p_1, p_2)$ to $\mathbf{p} = (a, b, c)$, and define

$$f(a, b, c) := a^2(1 + 2(b + c) + 6bc),$$

$$T(a, b, c) = \frac{a^2}{a^2 + b^2 + c^2} - f(a, b, c),$$

so that when $\mathbf{p} = (a, b, c)$, with a being the probability that a single sock has color 0, $T(a, b, c) = \mathbb{P}(X = 0) - \mathbb{P}(Y = 0)$. Note that $T(a, b, c) = T(a, c, b)$. Exchanging the roles among colors 0, 1, 2, we have $T(b, a, c) = \mathbb{P}(X = 1) - \mathbb{P}(Y = 1)$ and $T(c, a, b) = \mathbb{P}(X = 2) - \mathbb{P}(Y = 2)$. From Definitions 1 and 2, when $\mathbf{p} = (a, b, c)$,

$$2D(\mathbf{p}) = |T(a, b, c)| + |T(b, a, c)| + |T(c, a, b)|.$$

The expression above has the form $|T_1| + |T_2| + |T_3|$, and the absolute value function is an obstacle to taking the gradient. But by taking the eight cases for the sign, each of the expressions $\pm T_1 \pm T_2 \pm T_3$ is a rational function.

A straightforward parameterization of the two-dimensional set of probabilities (a, b, c) would have $a \geq b \geq 1 - a - b \geq 0$, implying that $T_1 \geq 0$ and $T_3 \leq 0$, so that there are only two cases, according to the sign of T_2. A major obstacle to this approach is the boundary, which is complicated, so instead we parameterize in terms of $(x, y) \in [0, 1]^2$ as follows:

$$\mathbf{p}(x, y) = (a, b, c) \text{ where } t = 1 + x + y, a = \frac{1}{t}, b = \frac{x}{t}, c = \frac{y}{t}.$$

Now taking $a = a(x, y)$ and so on, we have three functions defined on $[0, 1]^2$,

$$\begin{aligned} T_1(x, y) &:= T(a, b, c), \\ T_2(x, y) &:= T(b, a, c), \\ T_3(x, y) &:= T(c, a, b). \end{aligned}$$

The total variation distance is given by

$$2d_{\mathrm{TV}}(X, Y) = |T_1(x, y)| + |T_2(x, y)| + |T_3(x, y)|. \tag{12}$$

Since $1 \geq x, y$, we have $a \geq b, c$ and since the largest mass is at 1, we know that for all $x, y \in [0, 1]$, $T_1(x, y) \geq 0$.

We can eliminate the case $T_1 \geq 0, T_2 \geq 0$ and $T_3 \geq 0$, as this implies $T_1 = T_2 + T_3 = 0$ since $T_1 + T_2 + T_3 = 0$. By Lemma 2 this case gives $D(\mathbf{p}) = 0$, not of interest in the search for the maximum value. There are three remaining cases of sign to consider. Let

$$\begin{aligned} d_1(x, y) &= T_1(x, y) + T_2(x, y) - T_3(x, y), \\ d_2(x, y) &= T_1(x, y) - T_2(x, y) + T_3(x, y), \\ d_3(x, y) &= T_1(x, y) - T_2(x, y) - T_3(x, y). \end{aligned}$$

Then $\max d_{\mathrm{TV}}(X, Y) = \max(d_1, d_2, d_3)$, and so it suffices to check the maximum values of each of these rational functions.

Let us consider $g(x, y) := d_1(x, y)$.[6] Since g is a rational function in two variables, it is elementary to calculate the partial derivatives with respect to x and y, denoted g_x and g_y, respectively. What is *not* so elementary is finding all solutions (x, y) to the system $g_x(x, y) = g_y(x, y) = 0$. This set, $V(g_x, g_y) := \{(x, y) : g_x = g_y = 0\}$, also known as the affine *variety* defined by g_x, g_y, is what we wish to find; a good introductory text on this subject is [3].

[6]The term d_2 becomes d_1 under the interchange of x and y, so no further work is required for d_2. For d_3, the corresponding h_x and h_y, after cancellation of a common factor, have total degree 6 each, and one must account for the 36 solutions guaranteed by Bezout's Theorem.

Continuing with this example, even though g_x and g_y are rational functions, when each is rationalized it is clear that for $x, y \geq 0$ the denominator is always positive, and hence plays no role in characterizing the set of points in the variety $V(g_x, g_y) \cap [0,1]^2$. Thus we may simply find the variety of the numerators restricted to $[0,1]^2$, denoted h_x and h_y, respectively, which are bivariate polynomials.

A generalization to the Theorem of Algebra due to Bezout (see for example Chapter 5, Section 7 of [3]) can be used to verify that all solutions have been found[7]. In this case, after dividing out by a common factor of x, the two polynomials each have total degree 7. Bezout's theorem guarantees $7 \times 7 = 49$ solutions total including multiplicities, but some of these are solutions "at infinity."[8] Mathematica® finds a set of 19 unique, easily-verified solutions; when including multiplicities, this accounts for 39 of the total solutions. By hand we can find 10 solutions at infinity, so all 49 solutions have been addressed.

We obtain the largest value of d_{TV} from the point (x, y) given by[9]

$$\begin{aligned} x \in (0,1): \quad & 1 + 4x - 14x^2 - 4x^3 - 34x^4 + 20x^5 = 0, \\ y: \quad & y = x, \\ 2d_{\mathrm{TV}} = z \in (0, 0.2): \quad & 32000 + 168192z \\ & - 4557600z^2 + 14567472z^3 \\ & - 821583z^4 + 314928z^5 = 0. \end{aligned} \tag{13}$$

This solution is of the form

$$\mathbf{p} = \left(x_2, \frac{1 - x_2}{2}, \frac{1 - x_2}{2} \right)$$

[7]The precise form of the theorem requires several definitions and is not intended to be the focus; instead, we merely require assurance that the solutions found by Mathematica® [6] are exhaustive, since they are easily verified.

[8]Here is a simple analogy: How many times will a parabola intersect a line? A parabola has degree 2 and a line has degree 1. Suppose our parabola is $y = x^2$: then if our line is 1) $y = x - 1$, then there will be no intersections; 2) $y = 0$, then there is one intersection of multiplicity 2; 3) $y = x$, then there are two unique intersections of multiplicity 1 each; 4) $x = a$, for any real a, then there is one intersection of multiplicity 1. By using an appropriate transformation into the projective plane, one can guarantee exactly two solutions in all cases.

[9]The Mathematica® expressions are

$$x = \mathrm{Root}\left[1 + 4\#1 - 14\#1^2 - 4\#1^3 - 34\#1^4 + 20\#1^5 \ \&, 2\right],$$

$$d_{\mathrm{TV}} = \frac{1}{2}\mathrm{Root}[32000 + 168192\#1 - 4557600\#1^2 + 14567472\#1^3 - 821583\#1^4 + 314928\#1^5\&, 2].$$

for the value of $x_2 \in [0.5, 0.6]$ that solves $-5 + 42x_2 - 114x_2^2 + 168x_2^3 - 153x_2^4 + 54x_2^5 = 0$, with

$$x_2 \doteq 0.582011, \quad D(\mathbf{p}) \doteq 0.0842942; \tag{14}$$

the exact value of $D(\mathbf{p})$ given by Equation (13).

6. Conjectures about the largest possible discrepancy

The weakest conjecture is that there is some nontrivial upper bound on discrepancy. Formally, we define the universal constant for the pair discrepancy by

$$\ell_0 := \sup_{\mathbf{p}} D(\mathbf{p}), \tag{15}$$

where the supremum is over all distributions $\mathbf{p}$ on a finite or countable set of colors. Since total variation distance is always less than or equal to 1, trivially $\ell_0 \leq 1$, and the conjecture is

Conjecture 1. *The constant defined by* (15) *is* strictly *less than 1, i.e.,*

$$\ell_0 < 1. \tag{16}$$

6.1. **Conjectures for a finite number of colors.** If there are a finite number of colors, say $n+1$ with $n \geq 0$, then we can relabel the colors as $0, 1, \ldots, n$ so that $\mathbf{p} = (p_0, \ldots, p_n)$ with

$$p_0 \geq p_1 \geq \cdots \geq p_n \geq 0, \ p_0 + p_1 + \cdots + p_n = 1. \tag{17}$$

Given $n > 0$, and $x \in [\frac{1}{n+1}, 1)$, let

$$\mathbf{p}(n, x) = \left(x, \frac{1-x}{n}, \ldots, \frac{1-x}{n} \right), \tag{18}$$

which, due to $x \in [\frac{1}{n+1}, 1)$, satisfies (17).

With the notation (18), the result of Section 5.2 may be summarized as: for $n = 2$, over all probability distributions on $n+1$ colors standardized to satisfy (17), the maximum value of $D(\mathbf{p})$ is achieved, uniquely, at $\mathbf{p} = \mathbf{p}(2, x)$, with $x = x_2$ as specified by (14).

For each $n > 0$, (18) defines a *one parameter family* of probability distributions. At the endpoint $x = 1/(n+1)$, $\mathbf{p}(n, x)$ is a uniform distribution. Now suppose that $x \in (1/(n+1), 1)$, so that $\mathbf{p}(n, x)$ has $p_0 > p_1 = p_2 = \cdots = p_n > 0$. It is obvious from (2) that $\mathbb{P}(X = 0) > \mathbb{P}(X = 1) = \cdots = \mathbb{P}(X = n) > 0$, and Lemma 1 implies that $\mathbb{P}(Y = 0) > \mathbb{P}(Y = 1) = \cdots = \mathbb{P}(Y = n) > 0$. That is, both X and Y have distributions in the same one parameter family. Finally, (7) implies that $\mathbb{P}(X = 0) > \mathbb{P}(Y = 0)$, while for $i = 1$ to

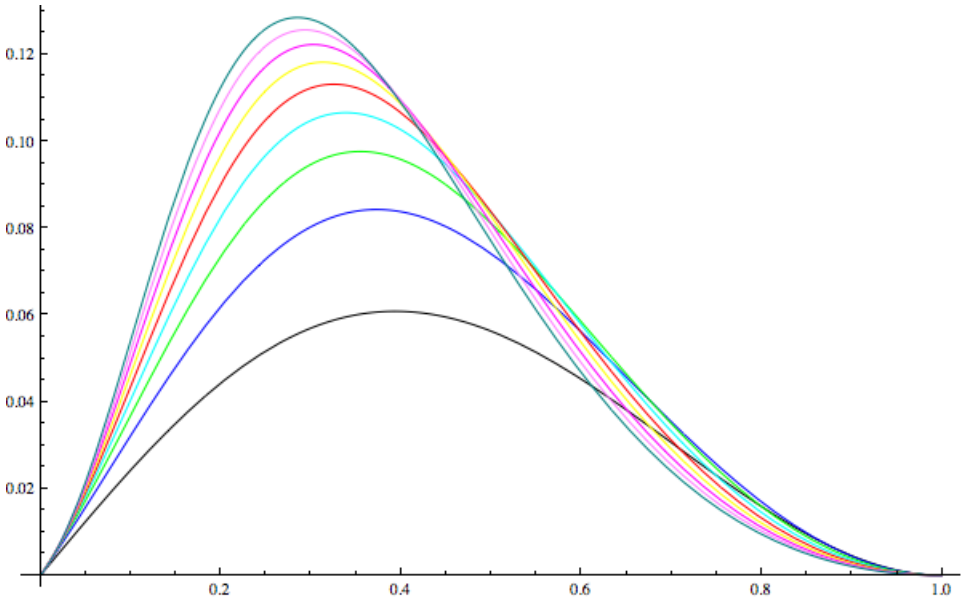

FIGURE 2. $D(\mathbf{p})$ for the one parameter families (18), $n = 1$ to 9. For each n, we plot $\frac{n+1}{n}x - \frac{1}{n}$ versus $D(\mathbf{p}(x,n))$, so that all 9 graphs have domain [0,1].

n, $\mathbb{P}(X = i) < \mathbb{P}(Y = i)$, and hence using (10), for each $n > 0$ and $x \in (\frac{1}{n+1}, 1)$, $\mathbf{p} = \mathbf{p}(n,x)$ has the simplified expression for its discrepancy,

$$\begin{aligned} D(\mathbf{p}) &= \mathbb{P}(X=0) - \mathbb{P}(Y=0) \\ &= \frac{x^2}{x^2 + \frac{(1-x)^2}{n}} - x^2 \sum_{k=0}^{n} (k+1)! \binom{n}{k} \left(\frac{1-x}{n}\right)^k . \end{aligned} \tag{19}$$

Conjecture 2. *For every nonnegative integer n, among all probability distributions on $n+1$ colors, the maximum value of $D(\mathbf{p})$ is achieved by a distribution of the form $\mathbf{p}(n, x_n)$.*

A slightly stronger conjecture is the following:

Conjecture 3. *For every nonnegative integer n, among all probability distributions on $n+1$ colors, the maximum value of $D(\mathbf{p})$ is achieved* uniquely *by $\mathbf{p}(n, x_n)$, where $x_n = \operatorname{argmax}_x D(\mathbf{p}(n,x))$.*

We cannot prove Conjecture 2, but we believe it to be true, for the following reasons.

(1) It is true, trivially for $n = 0$ and $n = 1$, and by Section 5.2, for $n = 2$.

(2) By broad analogy, many symmetric payoff functions achieve their extreme values at points with lots of symmetry. Indeed, Theorem 1 asserts that for each n, $D(\mathbf{p})$ achieves its *minimum* value, zero, at the uniform distribution, corresponding to the maximum conceivable symmetry in $\mathbf{p}$, while the family in (18) corresponds to *breaking* symmetry somewhat, but as little as possible.
(3) The one parameter family (18) shows up in other extremal problems which share the feature that the *labels* on the colors are irrelevant, and only the values of the probabilities matter. In particular, in information theory, the one parameter families show that "Fano's inequality is sharp;" see Cover and Thomas [2], (2.135) on page 40.
(4) For the moderate values $n = 3, 4, \ldots, 8$, when generating a million random points from the n-dimensional region specified by (17), the largest observed $D(\mathbf{p})$ in the sample came from a $\mathbf{p}$ that was close, by eye, to the form of (18).

The table below summarizes approximate extreme values under the one parameter families (18) for $n = 1, \ldots, 9$, using the notation $x_n = \text{argmax}_x D(\mathbf{p}(n, x))$.

$$\begin{array}{ll} x_1 = 0.6966599465951643196 & D(x_1) = 0.06084679923181354776 \\ x_2 = 0.5820110139097399105 & D(x_2) = 0.08429419234614604446 \\ x_3 = 0.5160030571683498864 & D(x_3) = 0.09766297359542326758 \\ x_4 = 0.4710812367633940106 & D(x_4) = 0.10661363736945495196 \\ x_5 = 0.4376598564845561514 & D(x_5) = 0.11316011048732238932 \\ x_6 = 0.4113811479448445739 & D(x_6) = 0.11822473613430355437 \\ x_7 = 0.3899258770101118464 & D(x_7) = 0.12229838762442936532 \\ x_8 = 0.3719239304877958135 & D(x_8) = 0.12566994796517442344 \\ x_9 = 0.3565033913388721410 & D(x_9) = 0.12852218802677888163 \end{array}$$

Figure 2 shows, for $n = 1$ to 9, $D(\mathbf{p}(x, n))$ for $x \in [\frac{1}{n+1}, 1]$; the graph plots $\frac{n+1}{n}x - \frac{1}{n}$ versus $D(\mathbf{p}(x, n))$, so that all 9 graphs use the same domain, [0,1].

7. Limit analysis of the one parameter family

Theorem 2. *For $c \in (0, \infty)$ define*

$$\ell(c) = \frac{c^2}{1 + c^2} - \int_0^\infty c^2 t e^{-ct - t^2/2}\, dt. \tag{20}$$

For any $c \in (0,\infty)$ and $n > 1/c^2$, let $\mathbf{p}^{(n)} = \mathbf{p}(n, c/\sqrt{n})$ be the distribution governed by (18) *with $x = c/\sqrt{n}$. Then*

$$\lim_{n\to\infty} D(\mathbf{p}^{(n)}) = \ell(c), \tag{21}$$

where ℓ is defined by (29)*.*

Proof. Extend Method 2 beyond the time of the first matching pair; i.e., pick socks forever. For each color i let N_i be the number of sock picks needed to get the second sock of color i. As the color varies, these random variables are *dependent*, since for any two distinct colors i, j and time $n \geq 2$, $0 = \mathbb{P}(N_i = N_j = n) < \mathbb{P}(N_i = n)\,\mathbb{P}(N_j = n)$. There is a standard technique to deal with this dependence, used in Markov chains[10], which is to take a sequence of independent exponentially distributed holding times $Y_1, Y_2, \ldots$, with $\mathbb{P}(Y_n > t) = e^{-t}$, and declare that the nth sock arrives at time $Y_1 + Y_2 + \cdots + Y_n$.[11] With values in $(0,\infty)$, the time T_i at which color i is first seen for the second time can be expressed as $T_i = Y_1 + \cdots + Y_{N_i}$. The distribution of the color of the first matching pair found, initially specified by (3), can also be expressed as

$$\mathbb{P}(Y = i) = P(T_i < \min_{j\neq i} T_j).$$

For each color i, the times at which socks of color i arrive form a Poisson arrivals process with rate p_i, and as the color varies, these processes are mutually independent; in particular the second arrival times T_i are mutually independent.

We are considering socks distributed according to $\mathbf{p}(n, c/\sqrt{n})$, that is, with $y := (1 - c/\sqrt{n})$,

$$p_0 = c/\sqrt{n}, p_1 = y/n, p_2 = y/n, \ldots, p_n = y/n. \tag{22}$$

Speed up time by a factor of $\sqrt{n}$; now socks of color 0 arrive at rate c, and for each other color $i = 1$ to n, socks of color i arrive at rate $p_i\sqrt{n} = y/\sqrt{n}$. For $t > 0$, and for each $i = 1$ to n, the number Z of socks of color i collected by time t is Poisson with parameter $\lambda = ty/\sqrt{n}$, and the event $\{T_i > t\}$ is the event $\{Z < 2\} = \{Z = 0 \text{ or } 1\}$, with

[10] see for example [5].

[11] The number of socks picked by time t is thus Poisson distributed, with mean t. Write $C_i(t) =$ the number of socks of color i chosen by time t. As i varies, the counts $C_i(t)$ are mutually independent; this observation is known as *Poissonization*. See exercise XII.6.3 in Feller [4].

probability

$$\begin{aligned}\mathbb{P}(T_i > t) &= \mathbb{P}(Z=0)+\mathbb{P}(Z=1)\\ &= e^{-\lambda}(1+\lambda)\\ &= \exp\left(-\frac{ty}{\sqrt{n}}\right)\left(1+\frac{ty}{\sqrt{n}}\right)\\ &= 1-\frac{t^2y^2}{2n}+O(n^{-3/2}).\end{aligned}$$

The easy way to see the result above is to argue that λ is small, so $e^{-\lambda}(1+\lambda) = (1-\lambda+\lambda^2/2-\lambda^3/6+\cdots)(1+\lambda) = 1-\lambda^2+\lambda^2/2+O(\lambda^3) = 1-\lambda^2/2+O(\lambda^3)$.

The event $\{\min(T_1,\ldots,T_n) > t\}$ is the intersection of the events $\{T_i > t\}$, so using the mutual independence, together with $y \to 1$,

$$\begin{aligned}\mathbb{P}(\min(T_1,\ldots,T_n) > t) &= \mathbb{P}(T_1 > t)^n = \left(1-\frac{t^2y^2}{2n}+O(n^{-3/2})\right)^n\\ &\to \exp(-t^2/2).\end{aligned}$$

Finally, we argue that the density of T_0, the second arrival time in a Poisson process with rate c, is given by

$$f(t) = c^2te^{-ct}.$$

This is a standard fact, known to some as the density of the Gamma distribution with shape parameter 2 and scale parameter c. Using the independence of T_0 and $\min(T_1,\ldots,T_n)$, we can condition on the value t for T_0 to get

$$\begin{aligned}\mathbb{P}_n(Y=0) &= \mathbb{P}(\min(T_1,\ldots,T_n) > T_0)\\ &= \int_0^\infty \mathbb{P}(\min(T_1,\ldots,T_n) > t) f(t)\ dt\\ &= \int_0^\infty \mathbb{P}(\min(T_1,\ldots,T_n) > t)\ c^2te^{-ct}\ dt\\ &\to \int_0^\infty c^2te^{-ct}e^{-t^2/2}dt.\end{aligned}$$

The above amounts to a calculation of the limit, as $n \to \infty$, of $\mathbb{P}_n(Y=0)$, corresponding to Method 2 when the underlying colors come from (22). For Method 1 the calculation is easier: using (1) we have $f_2 = p_0^2+p_1^2+\cdots+p_n^2 = (c/\sqrt{n})^2 + n(y/n)^2 = c^2/n + y^2/n$ and

$$\mathbb{P}_n(X=0) = \frac{p_0^2}{f_2} = \frac{c^2/n}{c^2/n+y^2/n} = \frac{c^2}{c^2+y^2} \to \frac{c^2}{c^2+1}.$$

At (19) we had already argued that once n is large enough that $p_0 > p_1$ we have the simplification, for our one parameter family, that $D(\mathbf{p}^{(n)}) = \mathbb{P}_n(X = 0) - \mathbb{P}_n(Y = 0)$. Combining this calculation of $D(\mathbf{p}^{(n)})$ with the limit values derived for $\mathbb{P}_n(Y = 0)$ and $\mathbb{P}_n(X = 0)$, (21) follows. □

We note that instead of invoking Poissonization, as in the above proof, one can argue directly with the explicit expression in (19), to show that under $x = c/\sqrt{n}$ and $k = t\sqrt{n}$, the sum in (19) is a Riemann approximation for $\int_0^\infty c^2 t e^{-ct} e^{-t^2/2} dt$.

8. Discussion

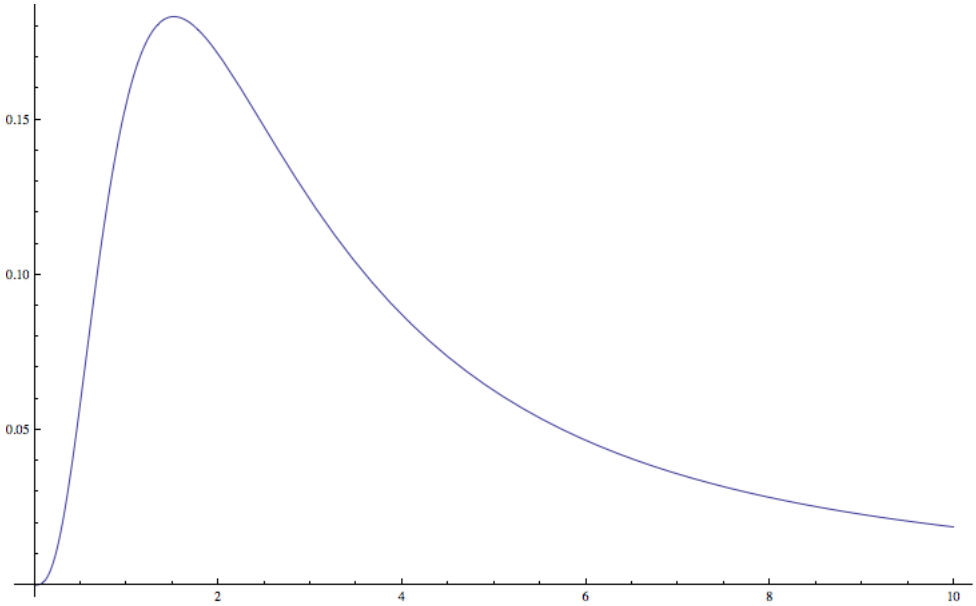

FIGURE 3. Plot of c versus $\ell(c)$ for $c = 0$ to 10. The maximum occurs at $c_0 \doteq 1.514$ and has value $\ell(c_0) \doteq 0.18320$.

If Conjecture 2 is true, it will follow that Conjecture 1 is also true, with the value of the universal constant for a pair of socks given by

$$\ell_0 = \sup_c \ell(c) = 0.1832000624087106\ldots \quad . \tag{23}$$

The argument requires two parts. The first part is to show that ℓ_0, defined in (15) as the sup of $D(\mathbf{p})$ over *all discrete* distributions, is equal to the sup over distributions with *finite* support. This is "soft" analysis, showing first that $\mathbf{p} \mapsto D(\mathbf{p})$ is continuous, hence given $\mathbf{p}$

with discrepancy greater than $\ell_0 - \varepsilon$ we can find a nearby distribution $\mathbf{p}'$ with finite support, close enough to $\mathbf{p}$ to guarantee that its discrepancy is greater than $\ell_0 - 2\varepsilon$. The second part, giving the concrete value for ℓ_0, uses compactness: given distributions $\mathbf{p}^{(n)} = \mathbf{p}(n, x_n)$ with discrepancies converging to ℓ_0, the values $c_n := x_n\sqrt{n} \in [0, \infty]$, $n \geq 1$, lie in a compact set, and hence there must be convergent subsequences. If $c_{n_k} \to c_0$ and $c_0 \in (0, \infty)$, then the proof of Theorem 2 already shows that the associated discrepancies converge to $\ell(c_0)$. If $c_{n_k} \to c_0$ with $c_0 = 0$ or $c_0 = \infty$, a small extension of the proof of Theorem 2 would show that the associated discrepancies would converge to 0. So indeed, $c_n \to c_0$ and $D(\mathbf{p}^{(n)}) \to \ell(c_0)$.

9. Shoes instead of socks: a matching left-right pair

Suppose, instead of wanting to collect a pair of matching socks, we want a pair of matching shoes. Naturally, this means one left shoe, and one right shoe, both of the same color. There are two reasonable ways to extend our study to this situation.

9.1. **One distribution for left colors, another distribution for right colors.** The setup here involves two discrete probability distributions, say $\mathbf{p}$ for the color S of a left shoe, and $\mathbf{q}$ for the color S' of a right shoe. The analog of (1) is

$$f_2 = \mathbb{P}(S = S') = \sum_i \mathbb{P}(S = S' = i) = \sum_i p_i\, q_i \tag{24}$$

for the probability that a random left shoe and a random right shoe match. We require that for at least one value i, $p_i q_i > 0$. The analog of (2) is the Method 1 distribution for the color $X = X(\mathbf{p}, \mathbf{q})$ of a matching left-right pair

$$\mathbb{P}(X = i) = \mathbb{P}(S = i | S = S') = \frac{p_i q_i}{f_2}. \tag{25}$$

For method 2, we assume that at times $1, 3, 5, \ldots$, one left shoe is collected, and at times $2, 4, 6, \ldots$, one right shoe is collected. Suppose that at time $k - 1$, there is not yet a matching left-right pair, but at time k, there is; then $Y = Y(\mathbf{p}, \mathbf{q})$ is the color of the shoe collected at time k.[12]

[12]There are other sensible ways to determine the matching color under sequential collection of shoes, for example, selecting one left and one right shoe each at time $1, 2, 3, \ldots$ and breaking ties via a coin flip. Even here, choices remain. For example, if the outcome is L_1 = red, R_1 = blue, L_2 = red, R_2 = white, L_3 = white, R_3 = red, then the tiebreak might be specified as equal odds for white versus red, or, since the available matches at time 3 are $(L_1, R_3), (L_2, R_3)$, and (L_3, R_2), as 2 to 1

The analog of discrepancy is now

$$D(\mathbf{p}, \mathbf{q}) = d_{\mathrm{TV}}(X(\mathbf{p}, \mathbf{q}), Y(\mathbf{p}, \mathbf{q})). \tag{26}$$

It is fairly easy to see that for this situation, the analog of Conjecture 1 is *false*; that is, the supremum of the discrepancy over all pairs of distributions is no smaller than the trivial upper bound on total variation distance:

$$1 = \sup_{\mathbf{p},\mathbf{q}} D(\mathbf{p}, \mathbf{q}). \tag{27}$$

We give a brief sketch of a proof of (27): with $a = a(n) = n^{-1/4}$ and $b = b(n) = n^{-2/3}$ let $\mathbf{p} = \mathbf{p}(n, a)$ and $\mathbf{q} = \mathbf{p}(n, b)$; in other words, $p_0 = \mathbb{P}(S = 0) = a$, $q_0 = \mathbb{P}(S' = 0) = b$ and for $i = 1$ to n, $p_i = \mathbb{P}(S = i) = (1 - a)/n$, $q_i = \mathbb{P}(S' = i) = (1 - b)/n$, with $a = n^{-1/4}$, $b = n^{-2/3}$. We have $p_0 q_0 = n^{-11/12}$ and

$$\sum_{i=1}^{n} p_i q_i = n \frac{1-a}{n} \frac{1-b}{n} \sim \frac{1}{n} = o(p_0 q_0),$$

so the Method 1 distribution converges to point mass at color 0, i.e., $\mathbb{P}_n(X = 0) \to 1$. To see that the Method 2 distribution has, in the limit, probability zero of getting color 0, consider collecting alternately left and right shoes forever. At time $m = 2n^{5/8}$, we will have collected $n^{5/8}$ left and $n^{5/8}$ right shoes. Thanks to the small value $q_0 = b = n^{-2/3}$, we expect only $n^{-1/24}$ left shoes of color 0 at time m, so with high probability, we do not yet have a matching pair of color 0. But, at time m, for *each* color $i = 1$ to n, the number of left shoes of color i is Binomial$(m, (1-a)/n)$, and hence is greater than zero with probability asymptotic to $m/n \sim n^{-3/8}$. Independently, the number of right shoes of color i is greater than zero with probability asymptotic to $n^{-3/8}$; hence the probability of at least one pair of color i is asymptotic to $n^{-3/4}$. The number W of colors $i > 0$ for which we have a pair has $\mathbb{E}\, W \sim n^{1/4}$, and the n events are negatively correlated with each other, so $\operatorname{Var} W < \mathbb{E}\, W$. By Chebyshev's inequality, $\mathbb{P}(W = 0) \leq \operatorname{Var} W/(\mathbb{E}\, W)^2 = O(n^{-1/4})$. So at time m, we are unlikely to have any pair of color 0, and unlikely not to have at least one pair of some other color, hence $\mathbb{P}_n(Y = 0) \to 0$.

---

in favor of red over white. For this outcome, our specification in the the main text is white, since the earliest match occurs at time 5, when $L_3 =$ white is observed.

9.2. **With the constraint $\mathbf{p} = \mathbf{q}$.** Now suppose that we declare that the distribution $\mathbf{p}$ for left shoes and the distribution $\mathbf{q}$ for right shoes must be equal. This does not reduce consideration of the distribution of a matching pair to the situation for socks; under the alternating left-right procedure, if we get a blue left shoe at time 1, a red right shoe at time 2, and another blue left shoe at time 3, then we still have not collected a matching pair.

The analog of Conjecture 1, for the situation of a matching left-right pair of shoes under the constraint of equal distributions, is plausible:

Conjecture 4.

$$\sup_{\mathbf{p}} D(\mathbf{p}, \mathbf{p}) < 1. \tag{28}$$

Furthermore, we can even propose a value for the universal constant for shoes, given by the left side of (28). It comes from an analog of Theorem 2. This analog of Theorem 2 is easiest to understand without the constraint $\mathbf{p} = \mathbf{q}$.

Theorem 3. *For $a, b \in (0, \infty)$ define*

$$\ell(a,b) = \frac{ab}{1+ab} - \int_0^\infty \left(ae^{-at} + be^{-bt} - (a+b)e^{-(a+b)t}\right) e^{-t^2}\, dt. \tag{29}$$

For $a, b > 0$ and sufficiently large n, let

$$\mathbf{p}^{(n)} = \mathbf{p}(n, a/\sqrt{n}),\ \mathbf{q}^{(n)} = \mathbf{q}(n, b/\sqrt{n}) \tag{30}$$

as in (18)*. Then*

$$\lim_{n\to\infty} D(\mathbf{p}^{(n)}, \mathbf{q}^{(n)}) = \ell(a,b). \tag{31}$$

Proof. The argument closely follows the proof for Theorem 2. We omit details, apart from sketching the main differences: under the distributions in (30), collecting left-right pairs with mean $1/\sqrt{n}$ holding times between pairs, the left shoes of color 0 form a rate a Poisson process, the right shoes of color 0 form a rate b Poisson process; $\mathbb{P}$(no left 0 by time t) $= e^{-at}$, $\mathbb{P}$(no right 0 by time t) $= e^{-bt}$, and *in the limit*, the two processes are independent, so $\mathbb{P}$(no left 0 and no right 0 by time t) $= e^{-(a+b)t}$. Inclusion-exclusion and differentiation leads to the limit density of the time T_0 at which a left-right pair of color 0 is found, $f(t) = \left(ae^{-at} + be^{-bt} - (a+b)e^{-(a+b)t}\right)$, instead of the c^2te^{-ct} of Theorem 2. At time t, for each of the n other colors we expect, asymptotically, $t/\sqrt{n}$ instances on the left, and $t/\sqrt{n}$ on the right, with t^2/n for the asymptotic chance of having a pair. This leads to $\mathbb{P}(\min(T_1, \ldots, T_n) > t) \to \exp(-t^2)$, instead of the $\exp(-t^2/2)$ of Theorem 2. □

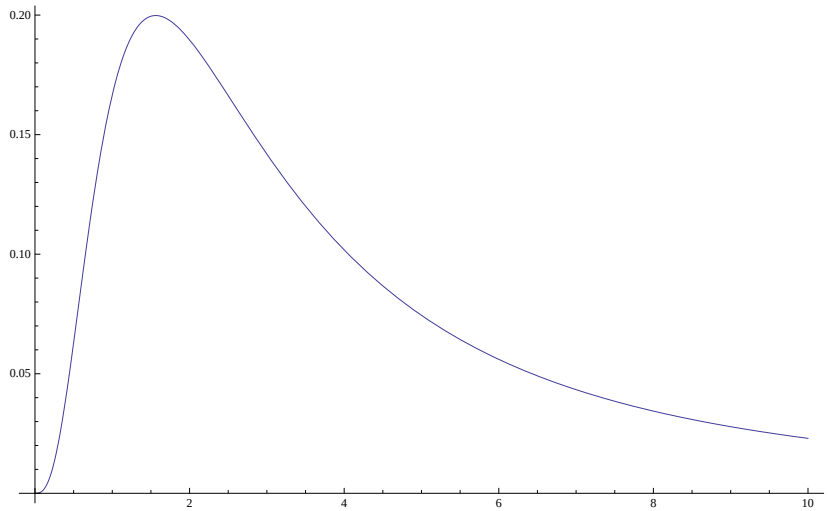

FIGURE 4. Plot of $\ell(a, a)$, the limit discrepancy $D(\mathbf{p}, \mathbf{q})$ when $\mathbf{p} = \mathbf{q} = \mathbf{p}(n, a/\sqrt{n})$. The maximum value $0.19980867\ldots$ occurs at $a = 1.562239\ldots$.

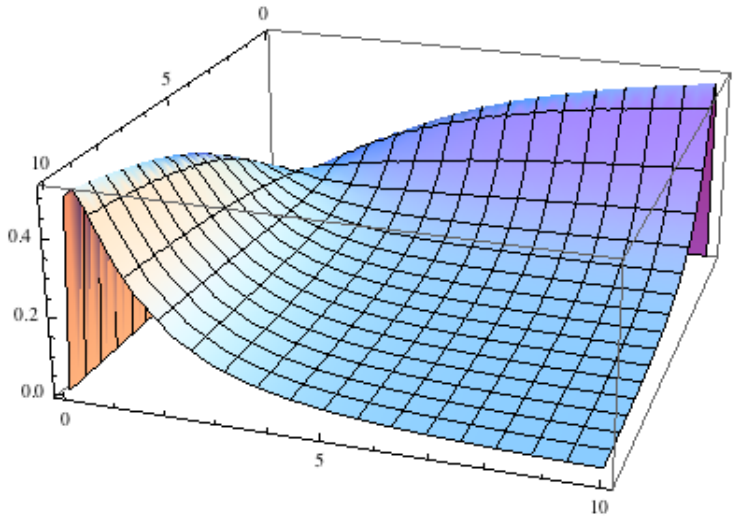

FIGURE 5. Plot of $\ell(a, b)$, the limit discrepancy $D(\mathbf{p}, \mathbf{q})$ when $\mathbf{p} = \mathbf{p}(n, a/\sqrt{n})$ and $\mathbf{q} = \mathbf{p}(n, b/\sqrt{n})$. The curve in Figure 4 lies along the diagonal, splitting the plot into two symmetric pieces.

While we do not have evidence for the analog of Conjecture 2 — indeed, it seems daunting to deal with the analog of Section 5.2, for left-right pairs under equal distribution for left and right — the analog

of Conjecture 1 *combined with* (23) is the following plausible conjecture. See Figure 4 for the source of the constant .1998 . . . .

Conjecture 5.

$$\sup_{\mathbf{p}} D(\mathbf{p}, \mathbf{p}) = \max_{a} \ell(a, a) \doteq 0.199808674053.$$

References

[1] Richard Arratia and Stephen DeSalvo. Probabilistic divide-and-conquer: a new exact simulation method, with integer partitions as an example. `arXiv:1110.3856v2 [math.PR]`, 2011.

[2] Thomas M. Cover and Joy Thomas. *Elements of Information Theory*. Wiley, 1991.

[3] David A. Cox, John Little, and Donal O'Shea. *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra, 3/e (Undergraduate Texts in Mathematics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[4] William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, January 1968.

[5] G.F. Lawler. *Introduction to Stochastic Processes*. Chapman & Hall Probability Series. Chapman & Hall, 1995.

[6] Mathematica. *Mathematica Edition: Version 8.0*. Wolfram Research, Inc., Champaign, IL, 2010.

(Richard Arratia) Department of Mathematics, University of Southern California, Los Angeles CA 90089.

E-mail address: `rarratia@math.usc.edu`

(Stephen DeSalvo) Department of Mathematics, UCLA Los Angeles CA 90095.

E-mail address: `stephendesalvo@math.ucla.edu`